CHEN BAI

Post-doctoral Fellow ⬦ Department of Electronic & Computer Engineering

Hong Kong University of Science and Technology, Hong Kong SAR

eecbai@ust.hk

`https://baichen318.github.io/`

RESEARCH INTERESTS

- Computer Architecture & Computer Systems

- Electronic Design Automation (EDA)

EDUCATION

The Chinese University of Hong Kong, NT, Hong Kong Aug. 2020 – Jul. 2024
Ph.D., Computer Science & Engineering
Supervisor: Prof. Bei Yu & Prof. Martin D.F. Wong

UESTC, Chengdu, P.R. China Sep. 2016 – Jun. 2020
B.Eng., Software Engineering
Provincial Outstanding Graduate

PUBLICATIONS

† indicates equal contribution; ‡ indicates the corresponding author.

Conference Papers

[C23] Ceyu Xu, Xiangfeng Sun, Weihang Li, **Chen Bai**, Bangyan Wang, Mengming Li, Zhiyao Xie, Yuan Xie, "PF-LLM: Large Language Model Hinted Hardware Prefetching", ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**), Pittsburgh, Mar. 22–26, 2026.

[C22] Xin Fan, **Chen Bai**‡, Xin Yang, Zhenhua Zhu, Yanhong Wang, Zhaode Liu, Yuan Xie, "FlashGEMM: Mesh-Aware Efficient GEMM for 3D-Stacked LLM Accelerators", IEEE/ACM Proceedings Design, Automation and Test in Europe (**DATE**), Verona, Apr. 20–22, 2026.

[C21] Qingchen Zhai, Hao Yu, **Chen Bai**, Charles Young, Frank Qu, Dezhi Ran, Yuan Xie, Tao Xie, "Towards Trustworthy LLM-Based Assertion Generation: A Data Augmentation Framework with Formal Check Approach", IEEE/ACM Proceedings Design, Automation and Test in Europe (**DATE**), Verona, Apr. 20–22, 2026.

[C20] **Chen Bai**†‡, Xin Fan†, Zhenhua Zhu, Wei Zhang, Yuan Xie, "AccelStack: A Cost-Driven Analysis of 3D-Stacked LLM Accelerators", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), Munich, Oct. 26–30, 2025.
(paper) (slides) (code)

[C19] Hongduo Liu, **Chen Bai**, Peng Xu, Lihao Yin, Xianzhi Yu, Hui-Ling Zhen, Mingxuan Yuan, Tsung-Yi Ho, Bei Yu, "LLMShare: Optimizing LLM Inference Serving with Hardware Architecture Exploration", ACM/IEEE Design Automation Conference (**DAC**), San Francisco, Jun. 22–25, 2025.
(paper) (slides) (poster)

[C18] Peng Xu, Su Zheng, Yuyang Ye, **Chen Bai**, Siyuan Xu, Hao Geng, Tsung-Yi Ho, Bei Yu, "RankTuner: When Design Tool Parameter Tuning Meets Preference Bayesian Optimization", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), New Jersey, Oct. 27–31, 2024.
(paper) (slides)

[C17] Yuanhang Gao, Donger Luo, **Chen Bai**, Bei Yu, Hao Geng, Qi Sun, Cheng Zhuo, "Is Vanilla Bayesian Optimization Enough for High-Dimensional Architecture Design Optimization?", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), New Jersey, Oct. 27–31, 2024.
(paper) (slides)

[C16] Lancheng Zou, Wenqian Zhao, Shuo Yin, **Chen Bai**, Qi Sun, Bei Yu, "BiE: Bi-Exponent Block Floating-Point for Large Language Models Quantization", International Conference on Machine Learning (**ICML**), Vienna, Jul. 21–27, 2024.
(paper) (slides) (poster)

[C15] Donger Luo, Qi Sun, Xinheng Li, **Chen Bai**, Bei Yu, Hao Geng, "Knowing The Spec to Explore The Design via Transformed Bayesian Optimization", ACM/IEEE Design Automation Conference (**DAC**), San Francisco, Jun. 23–27, 2024.
(paper) (slides)

[C14] Tong Qiao, Jianlei Yang, Yingjie Qi, Ao Zhou, **Chen Bai**, Bei Yu, Weisheng Zhao, Chunming Hu, "GN-Navigator: Towards Adaptive Training of Graph Neural Networks via Automatic Guideline Exploration", ACM/IEEE Design Automation Conference (**DAC**), San Francisco, Jun. 23–27, 2024.
(paper)

[C13] **Chen Bai**, Jianwang Zhai, Yuzhe Ma, Bei Yu, Martin D.F. Wong, "Towards Automated RISC-V Microarchitecture Design with Reinforcement Learning", AAAI Conference on Artificial Intelligence (**AAAI**), Vancouver, Feb. 20–27, 2024.
(paper) (slides) (code) (poster) (video)

[C12] Yuan Pu, Tinghuan Chen, Zhuolun He, **Chen Bai**, Haisheng Zheng, Yibo Lin, Bei Yu, "IncreMacro: Incremental Macro Placement Refinement", ACM International Symposium on Physical Design (**ISPD**), Taipei, Mar. 12–15, 2024.
(Best Paper Candidate)
(paper) (slides)

[C11] Shixin Chen, Su Zheng, **Chen Bai**, Wenqian Zhao, Shuo Yin, Yang Bai, Bei Yu, "SoC-Tuner: An Importance-guided Exploration Framework for DNN-targeting SoC Design", IEEE/ACM Asian and South Pacific Design Automation Conference (**ASPDAC**), South Korea, Jan. 22 – 25, 2024.
(paper) (slides)

[C10] **Chen Bai**, Xuechao Wei, Youwei Zhuo, Yi Cai, Hongzhong Zheng, Bei Yu, Yuan Xie, "Klotski: DNN Model Orchestration Framework for Dataflow Architecture Accelerators", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), San Francisco, Oct. 29 – Nov. 02, 2023.
(paper) (slides) (poster) (video)

[C9] Ziyang Yu, **Chen Bai**, Shoubo Hu, Ran Chen, Taohai He, Mingxuan Yuan, Bei Yu, Martin Wong, "IT-DSE: Invariant Risk Minimized Transfer Microarchitecture Design Space Exploration", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), San Francisco, Oct. 29 – Nov. 02, 2023.
(paper) (slides) (poster) (video)

[C8] **Chen Bai**, Jiayi Huang, Xuechao Wei, Yuzhe Ma, Sicheng Li, Hongzhong Zheng, Bei Yu, Yuan Xie, "ArchExplorer: Microarchitecture Exploration Via Bottleneck Analysis", IEEE/ACM International Symposium on Microarchitecture (**MICRO**), Toronto, Oct. 28 – Nov. 01, 2023.
(paper) (slides) (code) (poster)

[C7] **Chen Bai** †, Sicheng Li †, Xuechao Wei, Bizhao Shi, Yen-Kuang Chen, Yuan Xie, "2022 ICCAD CAD Contest Problem C: Microarchitecture Design Space Exploration", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), San Diego, Oct. 30 – Nov. 3, 2022. (**Invited Paper**)
(paper) (slides) (code)

[C6] Ziyi Wang, **Chen Bai**, Zhuolun He, Guangliang Zhang, Qiang Xu, Tsung-Yi Ho, Bei Yu, Yu Huang, "Functionality Matters in Netlist Representation Learning", ACM/IEEE Design Automation Conference (**DAC**), San Francisco, CA, Jul. 10 – 14, 2022.
(paper) (slides) (video)

[C5] Qi Sun, **Chen Bai**, Tinghuan Chen, Hao Geng, Xinyun Zhang, Yang Bai, Bei Yu, "Fast and Efficient DNN Deployment via Deep Gaussian Transfer Learning", IEEE International Conference on Computer Vision (**ICCV**), Oct. 11 – 17, 2021.
(paper) (appendix) (slides) (poster)

[C4] Zhuolun He, Ziyi Wang, **Chen Bai**, Haoyu Yang, Bei Yu, "Graph Learning-Based Arithmetic Block Identification", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), Nov. 1 – 4, 2021.
(paper) (slides)

[C3] **Chen Bai**, Qi Sun, Jianwang Zhai, Yuzhe Ma, Bei Yu, Martin D.F. Wong, "BOOM-Explorer: RISC-V BOOM Microarchitecture Design Space Exploration Framework", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), Nov. 1 – 4, 2021.
(William J. McCalla Best Paper Award)
(paper) (slides) (code) (video) (CUHK-CSE news) (ICCAD link)

[C2] Jianwang Zhai, **Chen Bai**, Binwu Zhu, Yici Cai, Qiang Zhou, Bei Yu, "McPAT-Calib: A Microarchitecture Power Modeling Framework for Modern CPUs", IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), Nov. 1 – 4, 2021.
(paper) (slides) (code) (video)

[C1] Qi Sun, **Chen Bai**, Hao Geng, Bei Yu, "Deep Neural Network Hardware Deployment Optimization via Advanced Active Learning", IEEE/ACM Proceedings Design, Automation and Test in Europe (**DATE**), Feb. 1 – 5, 2021.
(paper) (slides)

Journal papers

[J12] Peng Xu, Su Zheng, Yuyang Ye, **Chen Bai**, Siyuan Xu, Hao Geng, Tsung-Yi Ho, Bei Yu, "RankTuner: When Design Tool Parameter Tuning Meets Preference Bayesian Optimization", accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
(paper)

[J11] Lancheng Zou, Shuo Yin, Mingjun Li, Mingzi Wang, **Chen Bai**‡, Wenqian Zhao, Bei Yu, "Oiso: Outlier-isolated Data Format for Low-bit Large Language Model Quantization", accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
(paper)

[J10] Yuntao Lu, **Chen Bai**, Yuxuan Zhao, Ziyue Zheng, Yangdi Lyu, Mingyu Liu, Bei Yu, "DeepVerifier: Learning to Update Test Sequences for Coverage-Guided Verification", accepted by ACM Transactions on Design Automation of Electronic Systems (**TODAES**).
(paper)

[J9] Yuan Pu, Tinghuan Chen, Zhuolun He, Jiajun Qin, **Chen Bai**, Haisheng Zheng, Yibo Lin, Bei Yu, "IncreMacro: Incremental Macro Placement Refinement", accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
(paper)

[J8] Wenqian Zhao, Shuo Yin, **Chen Bai**, Zixiao Wang, Bei Yu, "BAQE: Backend-Adaptive DNN Deployment via Synchronous Bayesian Quantization and Hardware Configuration Exploration", accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
(paper)

[J7] **Chen Bai**, Xuechao Wei, Youwei Zhuo, Yi Cai, Hongzhong Zheng, Bei Yu, Yuan Xie, "Klotski v2: Improved DNN Model Orchestration Framework for Dataflow Architecture Accelerators", accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
(paper)

[J6] Ziyi Wang, **Chen Bai**, Zhuolun He, Guangliang Zhang, Qiang Xu, Tsung-Yi Ho, Yu Huang, Bei Yu, "FGNN2: A Powerful Pre-training Framework for Learning the Logic Functionality of Circuits", accepted by IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**).
(paper) (code)

[J5] Jianwang Zhai, Zichao Ling, **Chen Bai**, Kang Zhao, Bei Yu, "Machine Learning for Microarchitecture Power Modeling and Design Space Exploration: A Survey", Journal of Computer Research and Development (J-CRAD), vol. 61, no. 06, pp. 1-19, 2024. (in Chinese)
(paper)

[J4] **Chen Bai**, Qi Sun, Jianwang Zhai, Yuzhe Ma, Bei Yu, Martin D.F. Wong, "BOOM-Explorer: RISC-V BOOM Microarchitecture Design Space Exploration", ACM Transactions on Design Automation of Electronic Systems (**TODAES**), vol. 29, no. 01, pp. 1–23, 2024.
(paper)

[J3] Su Zheng, Hao Geng, **Chen Bai**, Bei Yu, Martin Wong, "Boosting VLSI Design Flow Parameter Tuning with Random Embedding and Multi-objective Trust-region Bayesian Optimization", ACM Transactions

on Design Automation of Electronic Systems (**TODAES**), vol. 28, no. 05, pp. 1 – 23, 2023.
(paper)

[J2] Ziyi Wang, Zhuolun He, **Chen Bai**, Haoyu Yang, Bei Yu, "Efficient Arithmetic Block Identification with Graph Learning and Network-flow", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**), vol. 42, no. 08, pp. 2591 – 2603, 2023.
(paper)

[J1] Jianwang Zhai, **Chen Bai**, Binwu Zhu, Yici Cai, Qiang Zhou, Bei Yu, "McPAT-Calib: A RISC-V BOOM Microarchitecture Power Modeling Framework", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**), vol. 42, no. 01, pp. 243 – 256, 2023.
(paper)

EXPERIENCE

Hong Kong University of Science and Technology, Hong Kong SAR   Feb. 2025 – present
Post-doctoral Fellow, Jockey Club STEM Lab of Future Advanced Computing Technologies
Advisor: Prof. Yuan Xie
Topic: 3D-stacked RISC-V-based computer architecture design & implementation
Description: My job focuses on the research for 3D-stacked RISC-V-based computer architecture analysis, design, and implementation for emerging applications.

Stealth Startup   Sep. 2024 – Jan. 2025
Advisory Researcher, Architecture Group
Description: Drive the implementation of research outcomes to achieve commercialization.

Alibaba DAMO Academy, Beijing, P.R. China   Jun. 2022 – May. 2024
Research Intern, Computing Technology Lab
Advisor: Prof. Yuan Xie, Dr. Hongzhong Zheng, and Dr. Xuechao Wei
Topic: Chip agile design methodology & Dataflow computing substrate
Description: My job focuses on the research for chip agile design methodology and some topics about dataflow computing substrate.

Huawei Hong Kong Research Center, Hong Kong SAR   Jun. 2021 – Apr. 2022
Research Intern, Turing Core & Key Technologies Development Department, HiSilicon HK
Advisor: Dr. Winnie Lo, and Dr. Xuanqi Chen
Topic: Microprocessor microarchitecture design space exploration & Power modeling
(Product: Huawei Mate70 Pro) (Product: Huawei Mate X6)
Description: My job focuses on research for the microarchitecture design space exploration and power modeling of the next-generation in-house microprocessor.

SenseTime, Beijing, P.R. China   Sep. 2019 – Jul. 2020
Research Intern, Intelligent Video Generation Group
Topic: SenseAR DigitalHuman – Audio-Driven Virtual Human (Product)(China Daily)
Advisor: Dr. Wayne Wu, and Dr. Qianyi Wu
Description: I work as a research intern, focusing on 2D and 3D digital human research and prototypes. The research and prototypes led to the commercial product SenseAR DigitalHuman, an audio-driven virtual human. The product has contributed over ten million RMB in profits to the SenseTime until the end of 2020.
(link) (China Daily)

Intel Asia-Pacific R. & D. Center, Shanghai, P.R. China   Feb. 2019 – Jul. 2019
Engineering Intern, Web Runtime Optimization Group
Topic: Chrome browser optimization for Intel architecture-based Chromebooks
Description: Conduct engineering implementation to accelerate Chrome brower applications for Intel architecture-based Chromebooks. Implemented codes are committed to the Chromium community. (code)

University of Maryland, Washington, D.C., U.S.A.   Jul. 2017 – Aug. 2017
Visiting Student
Topic: Study of "Leadership, Innovation, and Decision Making"

SELECTED AWARDS AND HONORS

William J. McCalla Best Paper Award	IEEE CEDA & ACM SIGDA	2021
ISPD Best Paper Candidate	IEEE CEDA & ACM SIGDA	2024
IEEE Circuits and Systems Student Travel Grant	IEEE CASS	2025
MICRO 2023 Student Travel Grants	TCuARCH, ACM SIGMICRO	2023
ICCAD Travel Support Grant	Futurewei, CEDA, and SIGDA	2023
Full Postgraduate Scholarship	The Chinese University of Hong Kong	2020 – 2024
Outstanding Graduate	The Education Department of Sichuan Province	2020
Excellent Thesis Award	UESTC	2020
National Scholarship	Ministry of Education	2017 – 2018
National Scholarship	Ministry of Education	2016 – 2017
Meritorious Winner of ICM	COMAP, INFORMS, SIAM, MAA, ASA, AMS	2018

TEACHING ASSISTANT

2023 Spring	CENG3420 Computer Organization and Design
2022 Spring	CENG3420 Computer Organization and Design
2021 Spring	CENG3420 Computer Organization and Design
2020 Fall	ENGG1110E Problem Solving By Programming

PROFESSIONAL SERVICE

Invited Talks

- 2025 Fudan-Guanghua International Forum for Young Scholars, Fudan University
  (link)

- 2023 Peisu XIA Forum, Institute of Computing Technology, Chinese Academy of Sciences
  (link)

- 2022 ACM SIGDA/IEEE CEDA Design Automation WebiNar (DAWN)
  (link) (video)

Journal Reviewer

- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)

- ACM Transactions on Design Automation of Electronic Systems (TODAES)

- IEEE Transactions On Very Large Scale Integration (VLSI) Systems (TVLSI)

- Integration, the VLSI Journal

- IEEE Design & Test

Conference Reviewer

- ACM/IEEE Design Automation Conference (DAC), 2022, 2023, 2024

- ACM/IEEE International Conference on Computer-Aided Design (ICCAD), 2022, 2023, 2024

- ACM/IEEE Asia and South Pacific Design Automation Conference (ASPDAC), 2022, 2023

- ACM/IEEE Workshop on Machine Learning for CAD (MLCAD), 2022

GRADUATE LEVEL COURSES

ENGG 5781	Matrix Analysis and Computations
ENGG 5301	Information Theory
CSCI 5160	Advanced Algorithms
CENG 5030	Energy Efficient Computing
SEEM 4340	Numerical Optimization
CSCI 5650	Graph Neural Networks
CSCI 5350	Advanced Topics in Game Theory
CENG 5270	EDA for Physical Design and Digital Systems

BASIC TECHNICAL SKILLS

Programming Languages	C/C++, Python, Java
Toolkits	Linux, MacOS, LaTeX

REFERENCES

Google Scholar

GitHub

LinkedIn